Journal of the Arkansas Academy of Science

Volume 53 Article 20

1999

Statistical Analysis of Climatic Variables in the Arkansas-Red River Basin

Felix Tendeku
University of Arkansas at Pine Bluff

Statistical Analysis of Climatic Variables in the Arkansas-Red River Basin

Felix Tendeku
Department of Mathematical Sciences and Technology
University of Arkansas at Pine Bluff
Pine Bluff, AR 71611

Abstract

Surface meteorological data for the Arkansas-Red River basin are analyzed in order to provide statistical data for modeling and simulation of climatic trends within the basin. The variables studied are the ambient temperature, temperature range, and precipitation. Daily and monthly mean values, spatial and seasonal variations, and frequency distributions are determined.

Introduction

This study is restricted to three surface meteorological variables, namely, ambient temperature, temperature range, and precipitation. These variables play an important role in climate change modeling, climatic impact studies and solar energy simulation studies. In agronomy, the study of climatic impact on crop growth and productivity requires data on daily global solar irradiation (H) for simulation models (Amir and Sinclair, 1991). Although historical global solar irradiation data are sparse, statistical relationships between H and other climatic variables provide a practical way to estimate H at locations where it was not measured (Barr et al., 1996; De Jong and Stewart, 1993). In solar heating and cooling studies, data on long wave atmospheric radiation are required. Often, as measured data are scarce or unavailable for the location under study, long wave radiation data may be estimated from meteorological data. Indeed, a large number of empirical formulae have been used for this purpose (Skartveit et al., 1996). Of the empirical formulae expressing irradiance in terms of surface-based variables, temperature and humidity dependent formulae have been applied with considerable success.

Historical data on precipitation amounts and spatial coverage are used in the design and operation of river-basin water-resource systems such as stream flow and reservoir regulation, real time operation of water works and hydrologic forecasting (Bras and Rodriguez-Iturbe, 1993).

The data used in this work was obtained from the University Consortium for Atmospheric Research (UCAR) Joint Office for Science Support (JOSS) Data Management System. It is archived as the National Climatic Data Center (NCDC) Summary of the Day Cooperative Data Set, one of several data sets provided for the Global Energy and Water-cycle Experiment (GEWEX) Continental-scale International Project (GCIP). The data contain surface meteorological observations from some 1450 stations within the Arkansas-Red River basin located approximately from

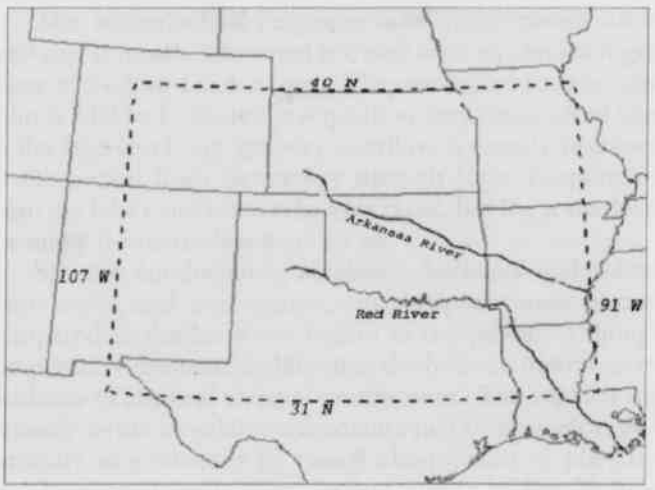

Fig. 1: A map showing Arkansas-Red River basin, the area under study, extending from 31N to 40N and 91W to 107W.

91W to 107W longitude and 31N to 40N latitude (Fig. 1). GEWEX has been initiated to study, on a global scale, the fast climate system mechanisms controlling radiation, cloud and rain, evaporation and fresh water storage. The Mississippi River basin provides a geographic area where significant atmospheric and hydrologic variations as well as land surface changes occur. The Arkansas-Red River basin, one of several intensively observed areas within the Mississippi River basin, represents different climatic and soil hydrology regimes to be studied during the GCIP project. Figure 2 shows the topography of the Arkansas-Red River basin. The time scales of the individual data sets used in this study are February 1 through April 30, 1992 (GIDS-1, 1992), April 1 through August 31, 1994 (GIST, 1994), April 1 through September 30, 1995 (GCIP/ESOP-95, 1995), and April 1 through September 30, 1996 (GCIP/ESOP-96,

1996). These data sets are compiled for the period of the year considered to have hydrological importance, that is, early spring through summer snowmelt and runoff.

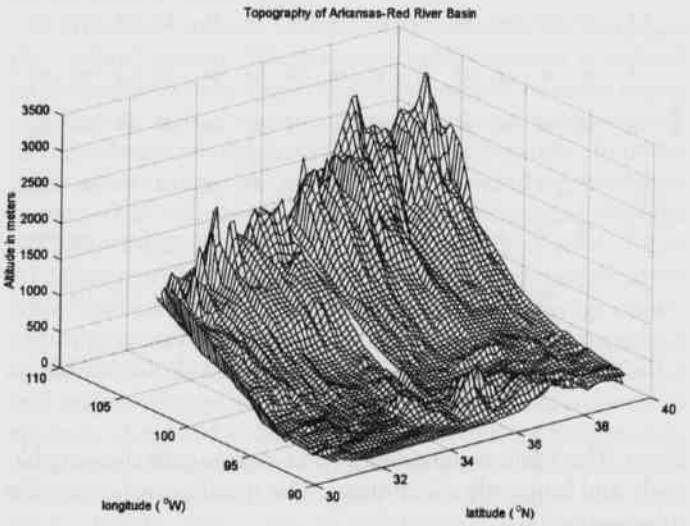

Fig. 2. Topography of Arkansas-Red River basin showing a westward rise in elevation from the Mississippi River (at approximately 91W longitude).

Data Analysis and Results

Time series data are formed for each variable by averaging daily readings over a spatial grid and a four-year period according to the following formula:

$$V_d = (1/NY)\sum_{y=1}^{Y}\sum_{n=1}^{N} V_{n,y} \qquad (1)$$

where $V_{n,y}$ is the value of the variable of the day for the nth station, y = 1,2,...,Y is the number of years of observation, n = 1, 2, ..., N is the number of stations with valid data points for a given day of the year, and V_d is the area-averaged daily value of the variable under consideration. Table 1 shows the number of data points per year used to determine V_d. The ambient temperature, T_a, and temperature range, T_r, are calculated from the daily minimum temperature, T_{min}, and maximum temperature, T_{max} as follows: $T_r = T_{max} - T_{min}$ and $T_a = T_{min} + T_r/2$. By merging the data sets using the above equation we obtain a time series for 243 days, from February 1 through September 30.

Figure 3 shows the time series of the area daily mean values of the climatic variables along with their daily spatial standard deviations, obtained by averaging data for the entire basin. These time series are analyzed to determine the monthly area statistics. A test of normality shows that all three variables deviate significantly from normal distribution. Since the underlying distribution is not precisely known, we employ a non-parametric method, the bootstrap

Year	Number of stations with data for:		
	Minimum Temperature	Maximum Temperature	Precipitation
1992	850	850	1435
1994	822	822	1334
1995	852	852	1470
1996	850	850	1382

Table 1. The number of stations providing data for the determination of basic statistics.

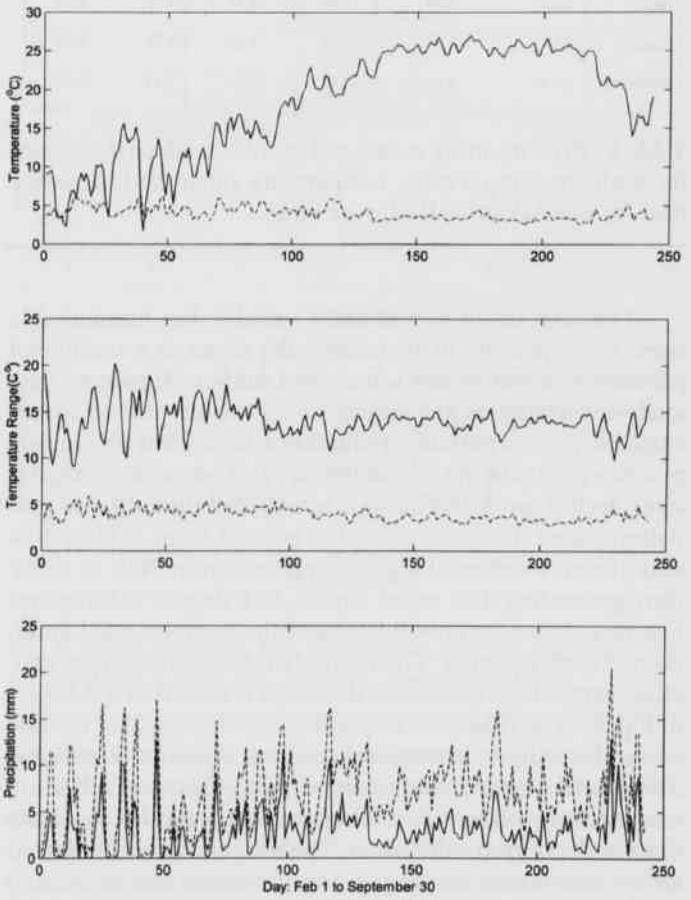

Fig. 3. A four-year averaged daily area mean (solid line) of climatic variables: air temperature, temperature range and precipitation with their spatial standard deviations (dashed line). Abscissa values refer to the day, from February 1 to September 30.

method (Efron and Tibshirani, 1993), to estimate the monthly mean and standard deviation. These statistics are reported in Table 2.

Month	Ambient Temperature °C		Temperature Range C°		Precipitation (mm)	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
February	7.26	4.57	13.67	4.30	2.03	4.91
March	10.49	4.58	14.94	4.37	2.21	4.70
April	13.82	4.57	15.50	4.45	2.74	5.90
May	18.99	4.47	13.22	4.23	3.82	8.21
June	33.99	3.55	13.95	3.64	2.96	7.66
July	25.65	3.58	13.94	3.51	3.17	8.36
August	25.43	3.24	13.64	3.29	2.84	7.66
September	20.12	3.64	13.41	3.66	3.41	7.86

Table 2. Area monthly mean and spatial standard deviations for ambient temperature, temperature range and precipitation for the Arkansas-Red river basin.

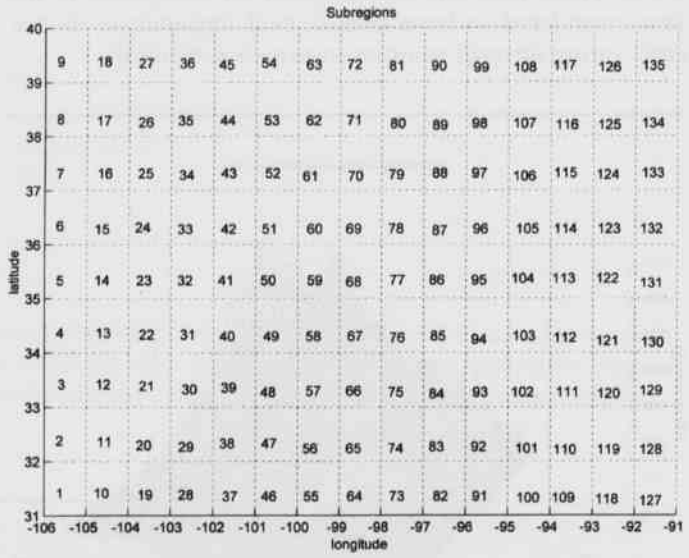

Fig. 4. The basin subdivided by 1x1 degree grid showing latitude and longitude coordinates. The numbers reference the sub-regions.

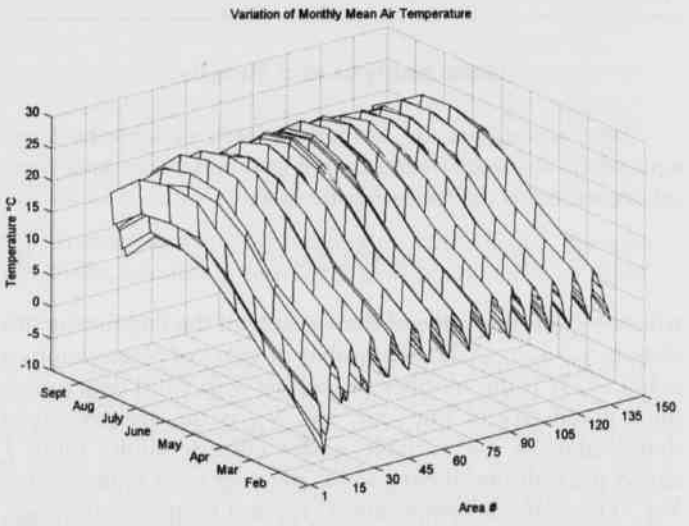

Fig. 5. Surface plot of air temperature versus location and month.

The area mean of a climatic variable has innumerable uses. For example, in hydrology the mean is a traditional parameter in runoff and water-yield studies. However, data analysis techniques are subject to uncertainty due to spatial variability. In a variance reduction scheme, it is a common practice to divide the region into several sub-areas of similar sizes. In this work the basin is subdivided into a 9 x 15 grid pattern, nine divisions along the latitude from 31N to 40N and fifteen divisions along the longitude from 91W to 106W thus generating 135 equal square 1x1 degree sub-regions. Figure 4 shows the coordinates and the numbers used to reference each sub-area. The monthly mean temperature variation versus location within the basin is shown in a 3-D plot in Fig. 5. The ridges correspond to areas along 31N latitude while the valleys correspond to areas along 40N latitude. The spatial trend is characterized by a positive north-south temperature gradient due to higher mean values in the south than in the north of the basin. The temporal trend is depicted by a nonlinear rise in the mean ambient temperature. It reaches a peak in July for areas in the south but in August for areas in the north.

After determining the basic statistics, attention is focused on the study of the time series data. Each time series may be considered to consist of two components:- (1) a steady periodic component composed of all seasonal cycles and (2) fluctuations which may be interpreted as the daily weather variations. The two components are studied separately and their characteristics are determined.

The Periodic Component.--To obtain the different seasonal cycles and their relative importance, the power spectrum is obtained for each area time series. It consists of finding the discrete-time Fourier transform of samples of the time series, taking the magnitude squared of the result, and scaling by the square of the norm of the data window applied to the time series (Welch, 1967). For ambient tem-

perature, the most significant component is 1 day/cycle. Other important harmonics occur at 2, 3, 4, 5, 8, and 11 days/cycle. For the temperature range, significant cyclical components occur at 1, 3, 4, 5, 12, and 13 days/cycle. Precipitation has significant seasonal components at 1, 4, 5, 22, 24, and 27 days/cycle.

Of special interest is the determination of the length of the "warm" season. Warm season in this context is defined as the period whose beginning in late winter or early spring and ending in late summer or early autumn are characterized by abrupt change in the ambient temperature. To determine when abrupt changes occur, a method of segmentation is used that is based on the Adaptive Forgetting through Multiple Model (AFMM) method (Ljung, 1994). The AFMM algorithm consists of building parallel models of the auto-regressive and auto-regressive moving-average types, assuming that the model parameters are piece-wise constant over time. It results in a model that splits the data record into segments over which the model remains constant. On applying the AFMM algorithm to the ambient temperature time series, abrupt changes in the ambient temperature were found to occur approximately on February 20 and September 23. The warm season is thus defined as the period from February 20 to September 23.

Residual Analysis.--After the significant cyclical components have been determined, they are subtracted from the original data, and the residuals thus formed are examined for statistical characteristics. If all cyclical components have been subtracted out, the residuals should have a zero mean. Thus we subject the residual signals to the test of hypothesis that the mean is zero. Using the sample mean, $\bar{y}$, the sample variance, S, the hypothesized population mean, μ, and the number of data points, N, then the statistic $t = (\bar{y} - \mu)\sqrt{N}/S$ will follow a Student's t-distribution with N- 1 degrees of freedom. For N = 243 if the test statistic is between ± 2.596 for $\alpha = 0.01$ level of significance, then the sample mean may be regarded as coming from a population with zero mean. Results of this test on the residual signals show the mean in each case is not significantly different from zero.

The residual signal for each climatic variable is separated into two temporal groups referred to as the spring and summer groups, respectively. A comparison is made to determine whether the frequency distribution is the same during the two periods. The Mann-Whitney statistic was used to test the null hypothesis that spring and summer samples follow the same frequency distribution. For both samples to be considered to have come from the same population, the Mann-Whitney statistic, M, must be between ± 1.96 at $\alpha = 0.05$ level of significance. Figure 6 shows the results of the Mann-Whitney statistic plotted for each of the three variables for all sub-regions within the basin. The test indicates that the frequency distribution of variables during spring may be considered to be the same as the summer distribution for all locations throughout the basin. In other words, the variables may be considered to have the same distribution throughout the warm season.

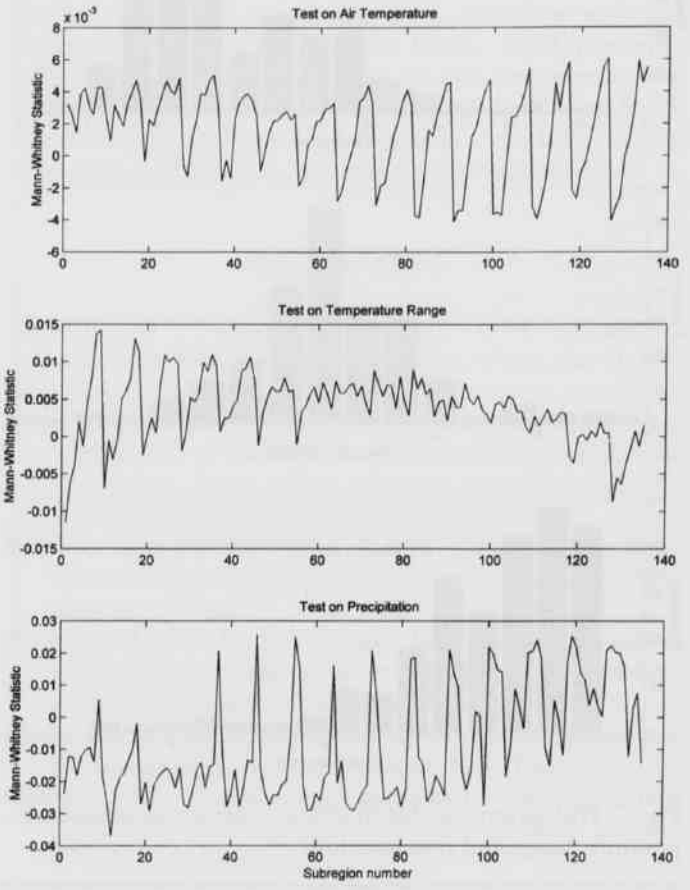

Fig. 6. Mann-Whitney test statistic for climatic variables versus locations within the Arkansas-Red River basin.

After examining the temporal distribution of variables, attention is focused on the examination of the spatial distribution of climatic variables. To test the uniformity of frequency distribution of variables between locations within the basin, the Kruskal-Wallis test was applied to see if samples from different locations could be considered to have come from the same population. If the size of samples from each population is at least 5 then the Kruskal-Wallis statistic, K, will closely follow a chi-square distribution (pchisq) with degrees of freedom, df = N-1, where N is the number of spatial samples. At a level of significance, α, the acceptance of the hypothesis that N samples follow the same distribution is given by $\text{pchisq}(K,df) < 1 - \alpha$. Results show that K is 0.82 for ambient temperature residuals, 1.96 for temperature

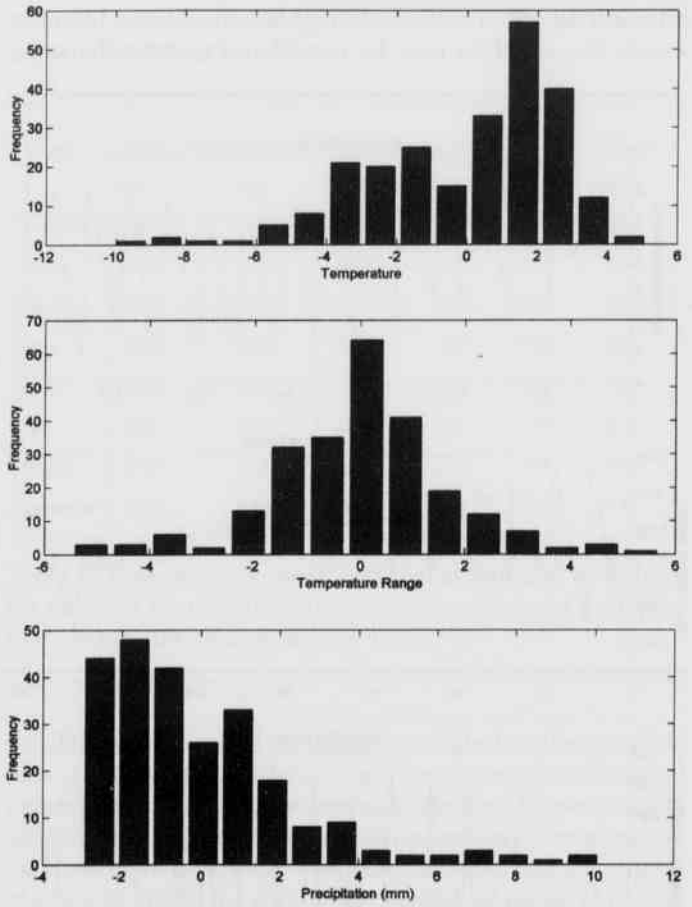

Fig. 7. Histograms of distributions of air temperature, temperature range and precipitation.

range residuals and 105.7 for precipitation residuals. Applying the Kruskal-Wallis test to the spatial samples gives values of pchisq(K,df) which are significantly less than 0.99 for each climatic variable. These results lead to the acceptance of the hypothesis that all variables follow the same distribution irrespective of location in the basin.

Frequency Distribution Functions.--The histograms of the residual signals (Fig. 7) are curve-fitted to a set of well known probability density functions, notably, the Gaussian, Lorentzian, Pearson VII, Log Normal, Symmetric and Asymmetric Double Sigmoid, Symmetric and Asymmetric Double Cumulative, Gamma, Weibull, Beta, Logistic, Pulse and Error functions. The choice of the best-fit function is based on the examination of three fit statistics, namely, the coefficient of determination, r^2, the fit standard error, S_e, and the F-statistic.

In the following, the mathematical relations of the fit statistics are given. The Sum of the Squares due to Error (SSE) is

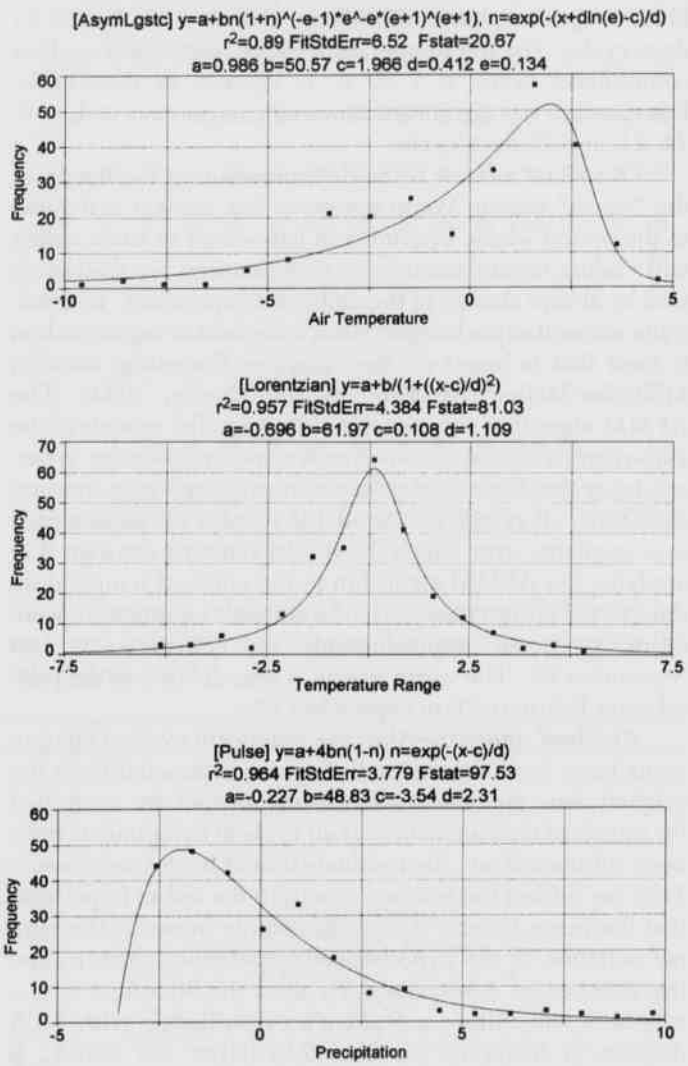

Fig. 8. Best-fit frequency distribution functions for air temperature, temperature range and precipitation.

$$SSE = \sum_{i=1}^{n} w_i(\hat{y}_i - y_i)^2 \qquad (2)$$

where SSE is the weighted (w) sum of the squared residuals, the data values are $\hat{y}_i$, the estimated values are $\hat{y}_i$, and the total number of data points is n. The Sum of the Squares about the Mean (SSM) is

$$SSM = \sum_{i=1}^{n} w_i(y_i - \bar{y})^2 \qquad (3)$$

where $\bar{y}$ is the arithmetic mean of the data values. The coefficient of determination is r^2=1-SSE/SSM. The degree of freedom, df, is defined as df = n-m, where m is the number of coefficients in the fit equation. The Mean Square Error, MSE, is defined as MSE = SSE/df. The Fit Standard Error

is $S_e = \sqrt{MSE}$. The Mean Square Regression, MSR = (SSM-SSE)/(m-1) and the F-statistic is defined as F = MSR/MSE.

The closer r^2 is to 1.0, the better the goodness of fit. However, this statistic is known to increase with increasing number of terms in an equation even though there may have been no real improvement in the fit. The Fit Standard Error, S_e, is the least-squares error of a fit. The closer this value is to zero the better the fit. It is directly related to the number of data points and is sensitive to outlying data points. The F-statistic is a measure of the extent to which a given equation represents the data. If an additional parameter makes a statistically significant contribution to a model, the F-statistic increases, otherwise it decreases. The higher the F-statistic the more efficiently a given equation fits the data. Figure 8 shows the best-fit frequency distribution functions for the residuals of the climatic variables. The Asymmetric Logistic function is the choice for the ambient temperature distribution, the Lorentzian function for temperature range and Pulse function for precipitation. The choice of the best-fit function is based on seeking the best compromise between r^2, S_e, and F-statistic.

Summary and Conclusions

Temperatures within the basin are time and location dependent. For all locations the temporal trend is characterized by a rise from February through September, reaching the peak in July for areas along 31N latitude while reaching the peak in August for areas along 40N latitude (Figs. 4 and 5). The peak values decrease northward, resulting in a north-south temperature gradient. The daily spatial variances are higher during spring than during summer. The frequency distribution during the warm season is found to follow approximately the Asymmetric Logistic function. Statistical tests show that the distribution is the same for all locations.

The daily temperature range, that is, the difference between daily maximum and minimum temperatures, exhibits the largest fluctuations during March and April. Peak values may attain 20C° while minimum values may drop to about 9C°. The warm season diurnal temperature range fluctuates between 12C° and 16C°. The temporal distribution, which is statistically the same for all locations, follows the Lorentzian function (Fig. 8).

Precipitation exhibits a very high degree of randomness with high spatial variances throughout the warm season. The highest four-year averaged monthly mean occurs in May. The frequency distribution follows the Pulse function (Fig. 8).

The work done so far has focused on the determination of the basic climatic statistics and general area trends within the basin. Our future work will include the determination of statistical models of both temporal and spatial variations, estimation of scales of spatial and temporal correlation, and the interdependence between the three surface meteorological variables considered in the present study.

ACKNOWLEDGMENTS.—The data sets used in this work were obtained from the University Corporation for Atmospheric Research (UCAR) Joint Office for Science Support (JOSS). I thank Dr. Tom DeFelice with the Atmospheric Science Group at the University of Wisconsin at Milwaukee for his collaboration on a related project using the GCIP data. The support received from Dr. Robert Jones, Associate Dean for Sponsored Programs at the University of Wisconsin at Milwaukee, is gratefully appreciated.

Literature Cited

Amir, J., and **T. R. Sinclair.** 1991. A model for temperature and solar radiation effects on spring wheat growth and yield. Field Crops Res. 28:47-58.

Barr, A. G., S. M. McGinn and **Si Beng Cheng.** 1996. A comparison of methods to estimate daily global solar irradiation from other climatic variables on the Canadian prairies. Solar Energy 56:213-224.

Bras, R. L., and **I. Rodriguez-Iturbe.** 1993. Random functions and hydrology. Dover Publications, Mineola, New York. xv+559pp.

De Jong, R., and **D. W. Stewart.** 1993. Estimating global solar radiation from common meteorological variables in western Canada. Can. J. Plant Sci. 73:509-518.

Efron, B., and **R. J. Tibshirani.** 1993. An Introduction to the bootstrap. Chapman and Hall, New York.

GIDS-1. 1992. Global energy and water-cycle experiment (GEWEX) Continental-scale international project (GCIP) Initial data set.

GIST. 1994. Global energy and water-cycle experiment (GEWEX) Continental-scale international project (GCIP) Integrated systems test.

GCIP/ESOP-95. 1995. Global energy and water-cycle experiment (GEWEX) Continental-scale international project (GCIP) 1995 Enhanced seasonal observing period.

GCIP/ESOP-96. 1996. Global energy and water-cycle experiment (GEWEX) Continental-scale international project (GCIP) 1996 Enhanced seasonal observing period.

Ljung, L. 1994. System identification - theory for the user. Prentice Hall, Englewood Cliffs, NJ.

Skartveit, A., J. A. Olseth, G. Czeplak and **M. Rommel.** 1996. On the estimation of atmospheric radiation from surface meteorological data. Solar Energy 56:349-359.

Welch, P. D. 1967. The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE. Trans. Audio Electroacoust. AU-15: 70-73.